IST 718:
Big Data Analytics

UNIT 1-1 Math Review

Note: MATERIAL here is from a mix of Readings & sources as listed in References Slide(s)

Math Review

Linear Algebra

Calculus

Probability

Statistics

Linear Algebra

Linear Algebra Introduction

Linear algebra is the branch of mathematics that deals with the theory of systems of linear equations, matrices, vector spaces, and linear transformations. Quote by Gilbert Strang (MIT): "Linear Algebra has become as basic and applicable as Calculus, and fortunately it is easier."

Linear Algebra appears in virtually every branch of applied mathematics, physics, statistics, mathematical economics, etc. – and is very important to:

- Big Data Analytics
- Machine learning – especially Deep Learning
- Inferential and Exploratory Statistics
- Many other fields in Science

Here, we take an "application oriented" approach to Linear Algebra:

- Focused on Linear Algebra subset most relevant to Machine Learning

Vectors

Scalars

- Represented by Greek letters α,β,γ
- Represents integers, reals, rationals, etc.
- Scalars are quantities that are fully described by a magnitude alone.
- Examples:
  α=0.1
  β=1^(−10)
  γ = 3

Vectors

- Represented by lower case letters
- A vector is a 1-D array of scalars:
- $a=\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_1 \end{bmatrix}$
- A vector is also a single column matrix
- In data science, a vector is often just a column of numbers that describes an event.

Matrices AND Tensors

Matrix (plural, Matrices)

- A 2-D array of scalars
- Represented by capital letters like A, B, C, etc.
- The matrix dimension is denoted $A\_(m\times n)$ where m = number of rows and n = number of columns
- $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$
- The above matrix has m rows and n columns

Tensor

- A tensor is a generalized matrix
- A tensor includes all of the following
  - 0 dimension (AKA scalar)
  - 1 dimension (AKA vector)
  - 2 dimension (AKA matrix)
  - Or higher dimension
- A tensor is more general and flexible than a matrix
- The dimension of the tensor is known as it's "rank"

Data representation

Data Forms

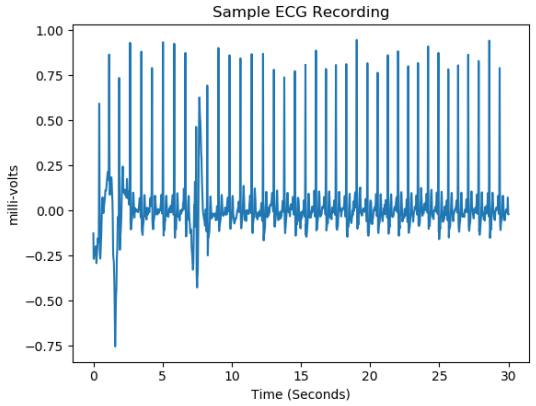
- Text (regular)
- Geometrical shapes
- Graphs
- Networks
- Sequences
- Tables
- Multimedia
  - Images
  - Sound
  - Video

Common Data Formats

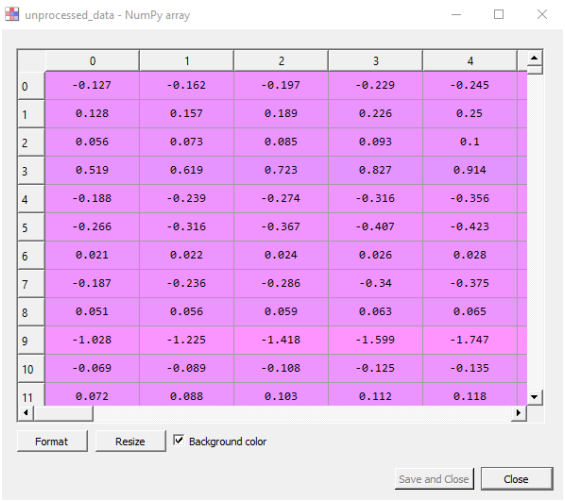
- Text file
- Tensor data
- Image data
- Video data
- Timeseries data
- Comma-separated Text
- Parquet files

Data Example

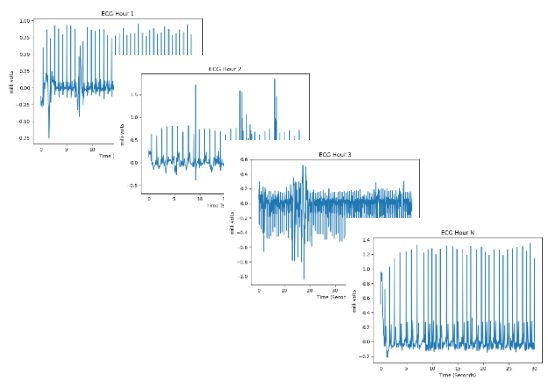
EKG Data Vector



EKG Matrix: A Col is A Recording



	0	1	2	3	4
0	-0.127	-0.162	-0.197	-0.229	-0.245
1	0.128	0.157	0.189	0.226	0.25
2	0.056	0.073	0.085	0.093	0.1
3	0.519	0.619	0.723	0.827	0.914
4	-0.188	-0.239	-0.274	-0.316	-0.356
5	-0.266	-0.316	-0.367	-0.407	-0.423
6	0.021	0.022	0.024	0.026	0.028
7	-0.187	-0.236	-0.286	-0.34	-0.375
8	0.051	0.056	0.059	0.063	0.065
9	-1.028	-1.225	-1.418	-1.599	-1.747
10	-0.069	-0.089	-0.108	-0.125	-0.135
11	0.072	0.088	0.103	0.112	0.118



Norms

Norm properties

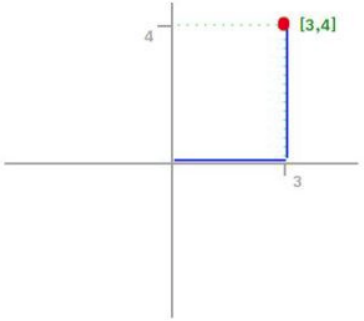
- A norm is a function that measures how “large” a vector is.
- Similar to the distance between zero and the point represented by the vector.
- Often referred to as L^P Norm where p is the order of norm function
- $\| x \|_p = (\sum_i |x_i|^p)^{1/p}$

- $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
- $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (the *triangle inequality*)
- $\forall \alpha \in \mathbb{R}, f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$

Measuring L^1 Norm and L^2 Norm

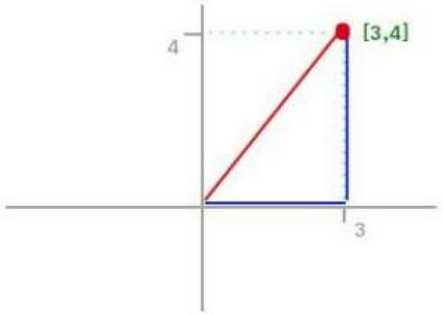
L^1 Norm

- is sometimes called the taxicab or Manhattan norm because it represents the total distance traveled between the start and end of the vector
- is one of the most common Norms
- $\| x \|_1 = |3| + |4| = 7$



L^2 Norm

- Calculates the shortest distance between the start
- Also known as the "Euclidian" Norm
- $\| x \|_2 = \sqrt{3^2 + 4^2} = 5$



Vector Operations

Inner (Dot) Product

- The dot product of 2 vectors yields a scalar
- $a * b = \sum_i a_i b_i$
- $[a_1 \;\; a_2 \;\; a_3] * [b_1 \;\; b_2 \;\; b_3] = \; [a_1 b_1 + a_2 b_2 + a_3 b_3]$

Outer (Cross) Product

- The cross product of 2 vectors yields a matrix
- $u \otimes v = uv^t$
- $$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} [v_1 \; v_2 \; v_3] = \begin{bmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 \\ u_2 v_1 & u_2 v_2 & u_2 v_3 \\ u_3 v_1 & u_3 v_2 & u_3 v_3 \\ u_4 v_1 & u_4 v_2 & u_4 v_3 \end{bmatrix}$$

Matrix Operations – Add / Subtract / Transpose

- Matrices must be the same size to add or subtract
- Matrix addition is commutative: A + B = B + A
- **Add** or **Subtract** each matrix element between operands
- Resulting matrix has the same size as the 2 matrix operands
- A + B = C → $\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$
- **Transpose** will swap rows and columns: $(A^T)_{i,j} = A_{j,i}$
- $A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix}$ $A^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$
- Property: $(AB)^T = B^T A^T$

Matrix Multiplication

- To **Multiply a matrix by a scalar**, "broadcast" the scalar across all of the matrix elements and multiply:
  - $\mathrm{A} * 2 = \mathrm{C} \rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} * 2 = \begin{bmatrix} 2a & 2b \\ 2c & 2d \end{bmatrix}$
- To **Multiply a matrix by a matrix**, the number of columns in A must match the number of rows in B: multiply each row element of A by the corresponding column element in B and sum the result:
  - $\sum_k A_{i,k} B_{k,j} = C_{i,j}$
  - $\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} * \begin{bmatrix} q & r & s & t \\ u & v & w & x \end{bmatrix} = \begin{bmatrix} (a*q+b*u) & (a*r+b*v) & (a*s+b*w) & (a*t+b*x) \\ (c*q+d*u) & (c*r+d*v) & (c*s+d*w) & (c*t+d*x) \\ (e*q+f*u) & (e*r+f*v) & (e*s+f*w) & (e*t+f*x) \end{bmatrix}$
  - Matrix multiplication is not commutative: AB != BA
  - The size of the resulting matrix C is equal to the number of rows in A and the number of columns in B

Identity Matrix and Matrix Inversion

Identity Matrix

- I is the identity matrix when:
  - A*I = A
- Multiplying a matrix by the identity matrix produces the original matrix.
- Identity matrix has ones on its diagonal and zeros elsewhere
  - Example: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Inverse Matrix

- A^{-1} is the inverse matrix when:
  - $A^{-1}A = I$
  - for some matrix A
- I is the identity matrix
- Goodfellow notes that inversion is numerically unstable but useful for abstract analysis
  - Due to the numeric instability, the computer uses different methods to solve

- **Matrices are not invertible if**:
  - More rows than columns OR more columns than rows (not square)
  - There are linearly dependent rows / columns
    - Example: one row or col is a multiple of another row or col
- A square matrix that is not invertible is called “singular” or “degenerate”

Linear Combinations

- A **linear combination** is an expression constructed from a set of terms by multiplying each term by a constant and adding the results.
  - Example: The linear combination of the terms x and y would be constructed by multiplying each of the terms by a constant and adding.
  - ax + by where a and b are constants
- A linear combination can be performed on the columns or rows of a matrix
- Linear combinations are very important to **principal component analysis**
- Example matrix column linear combination for $\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \end{bmatrix}$
- $c_1 \begin{bmatrix} u_{11} \\ u_{21} \\ u_{31} \\ u_{41} \end{bmatrix} + c_2 \begin{bmatrix} u_{12} \\ u_{22} \\ u_{32} \\ u_{42} \end{bmatrix}$

Systems of Equations

- The ':' operator means all rows or columns. Examples:
  - Subscript '1,:' means row 1, all columns
  - Subscript ':,2' means all rows of column 2
- The equation Ax=b expands to:
  - Row 1 of the matrix equals the first answer: $A_{1,:} * x = b_1$
  - Row 2 of the matrix equals the second answer: $A_{2,:} * y = b_2$
  - Row n of the matrix equals the nth answer: $A_{n,:} * z = bn$
- A linear system of equations may have: no solution, many solutions, or exactly one solution which means multiplication by the matrix is an invertible function

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{2.22}$$

$$\boldsymbol{A}^{-1}\boldsymbol{Ax} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.23}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.24}$$

Calculus

What is a limit?

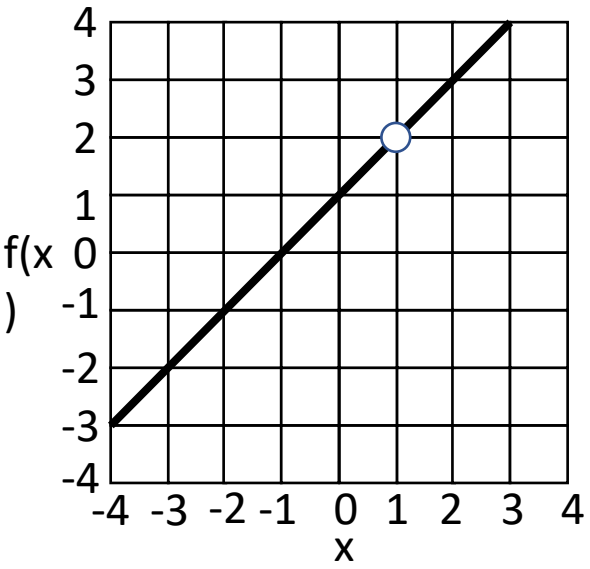
- A limit is what happens when you get closer and closer to a point without actually reaching it.
- Example: If $f(x) = 2x$ then as $x \to 1$, $f(x) \to 2$.
- We write this as $\lim_{x \to 1} f(x) = 2$.

x	0	.9	.99	.999	.9999
f(x)	0	1.8	1.98	1.998	1.9998

Why are limits useful?

- Many functions are not defined at a point but are well-behaved nearby.
- Example: If $f(x) = \frac{x^2-1}{x-1}$ then $f(1)$ is undefined.
- However, as $x \to 1$, $f(x) \to 2$, so $\lim_{x \to 1} f(x) = 2$.

x	0	.9	.99	.999	.9999
f(x)	0	1.9	1.99	1.999	1.9999



Left Limits and Right Limits

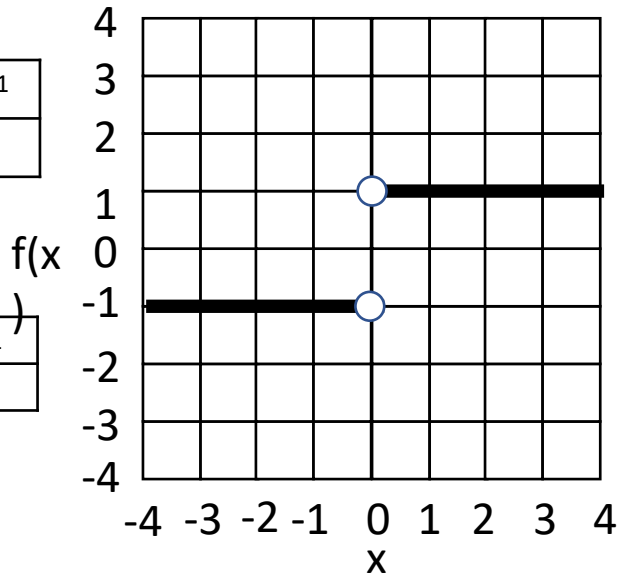
Consider $f(x) = \frac{x}{|x|}$. $f(0)$ is undefined.

As $x \to 0^-$, $f(x) = -1$

x	-1	-.1	-.01	-.001	-.0001
f(x)	-1	-1	-1	-1	-1

As $x \to 0^+$, $f(x) = 1$

x	1	.1	.01	.001	.0001
f(x)	1	1	1	1	1

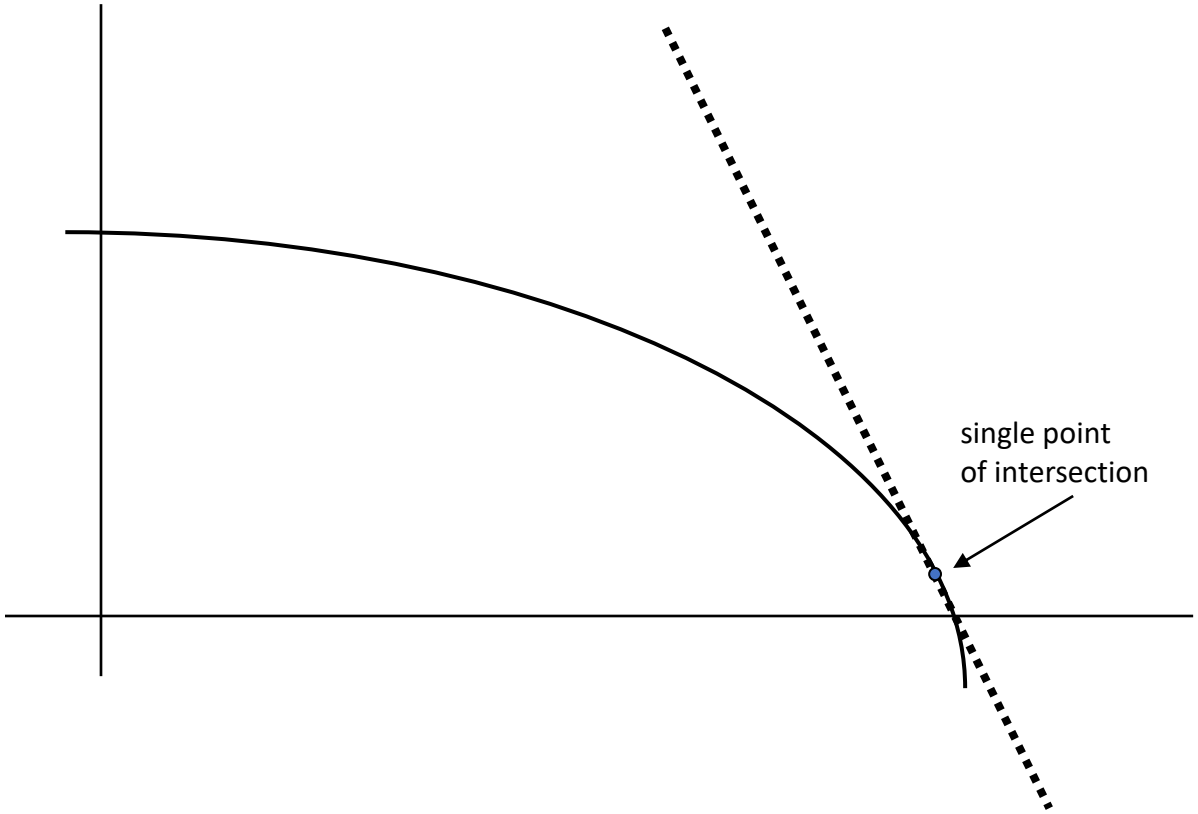


We write this as $\lim_{x \to 0^-} f(x) = -1$, $\lim_{x \to 0^+} f(x) = 1$

What is a *derivative*?

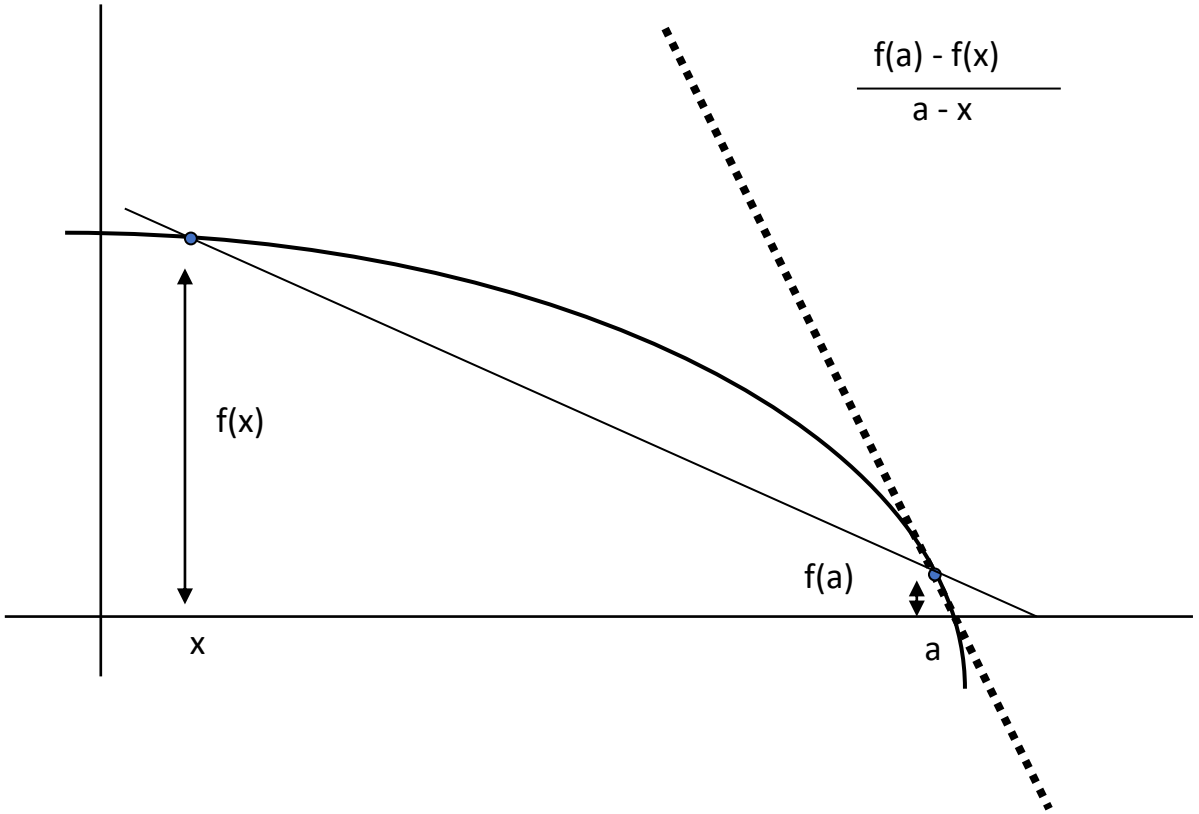
- Given a function
- A derivative is the rate of change of that function
- Can be represented by slope of the line tangent to the curve

The tangent line

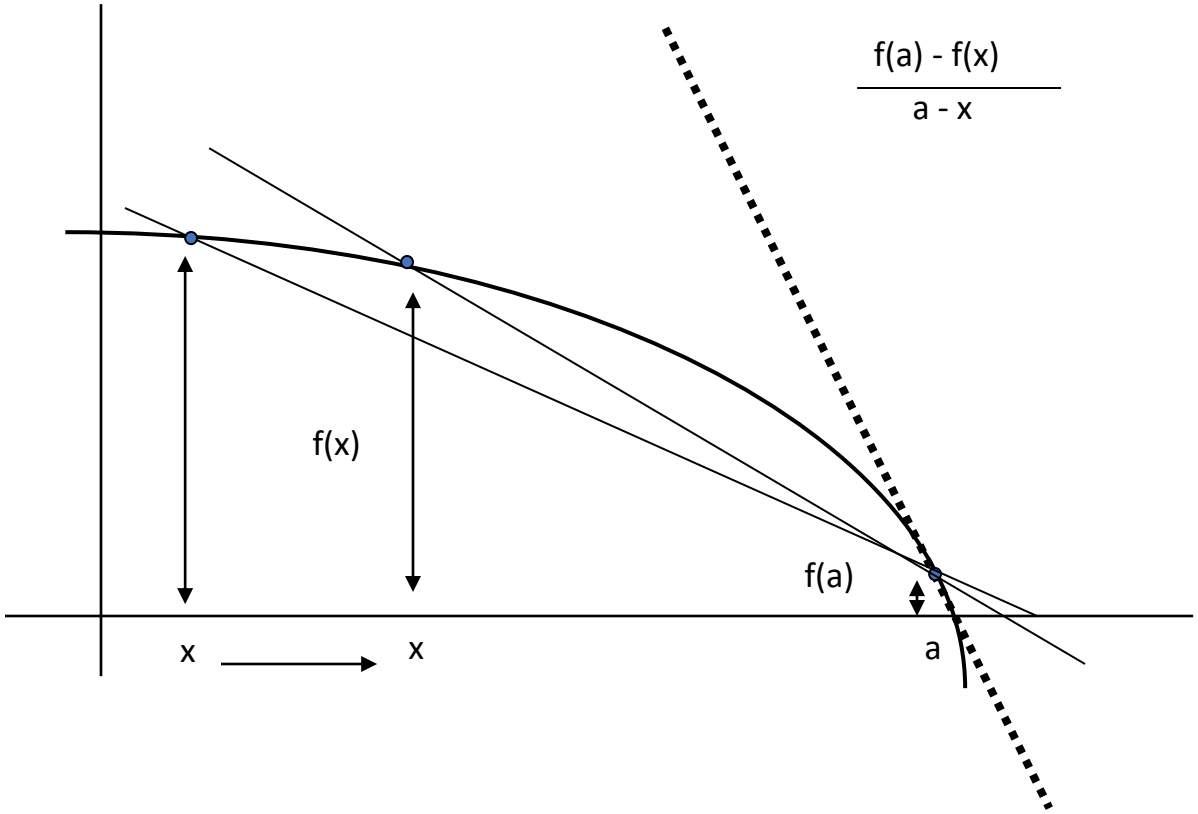


slope of a secant line

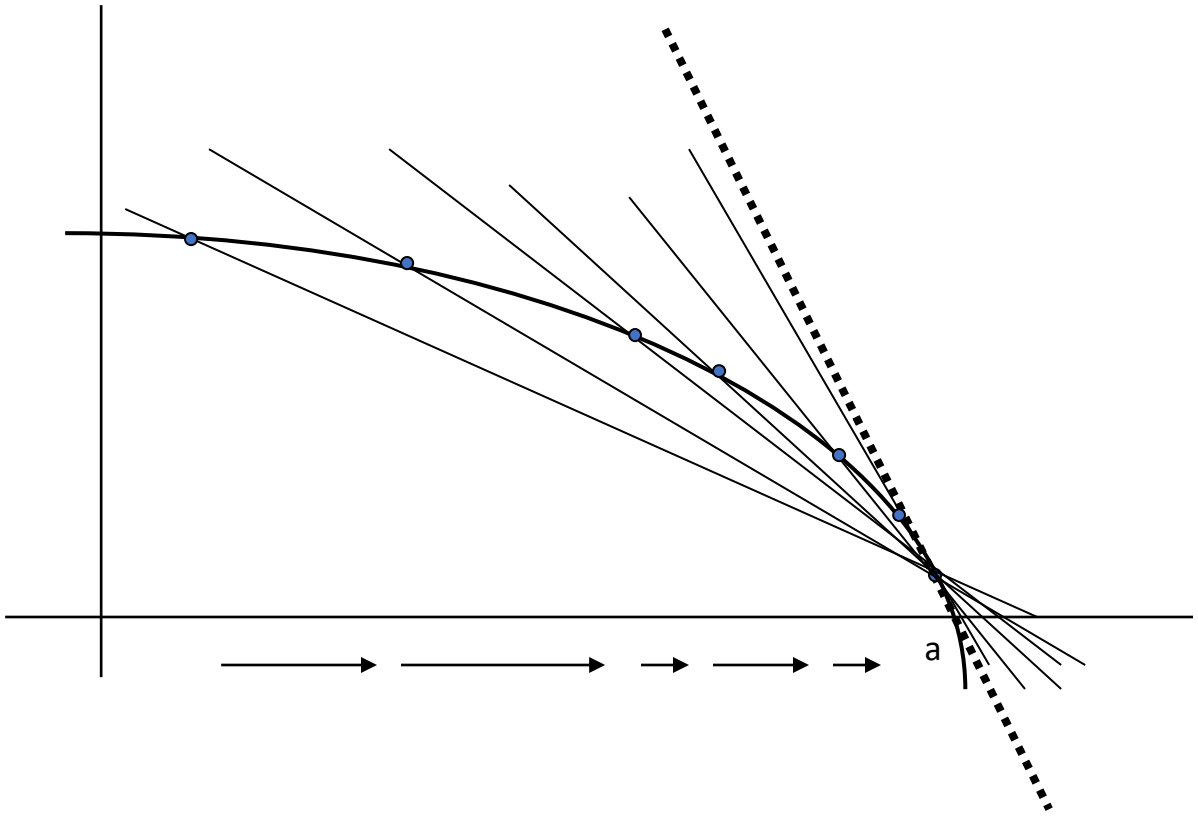
- A secant line connects 2 points on a curve together



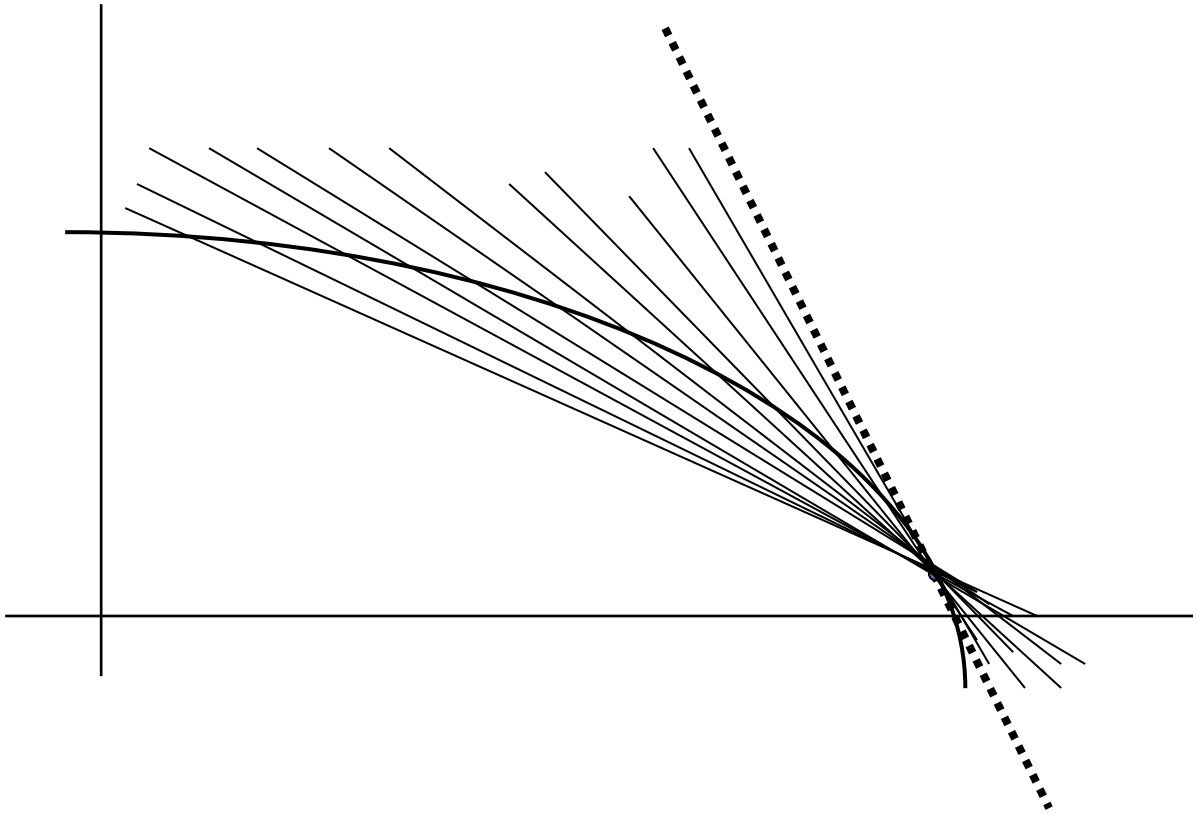
slope of a (closer) secant line



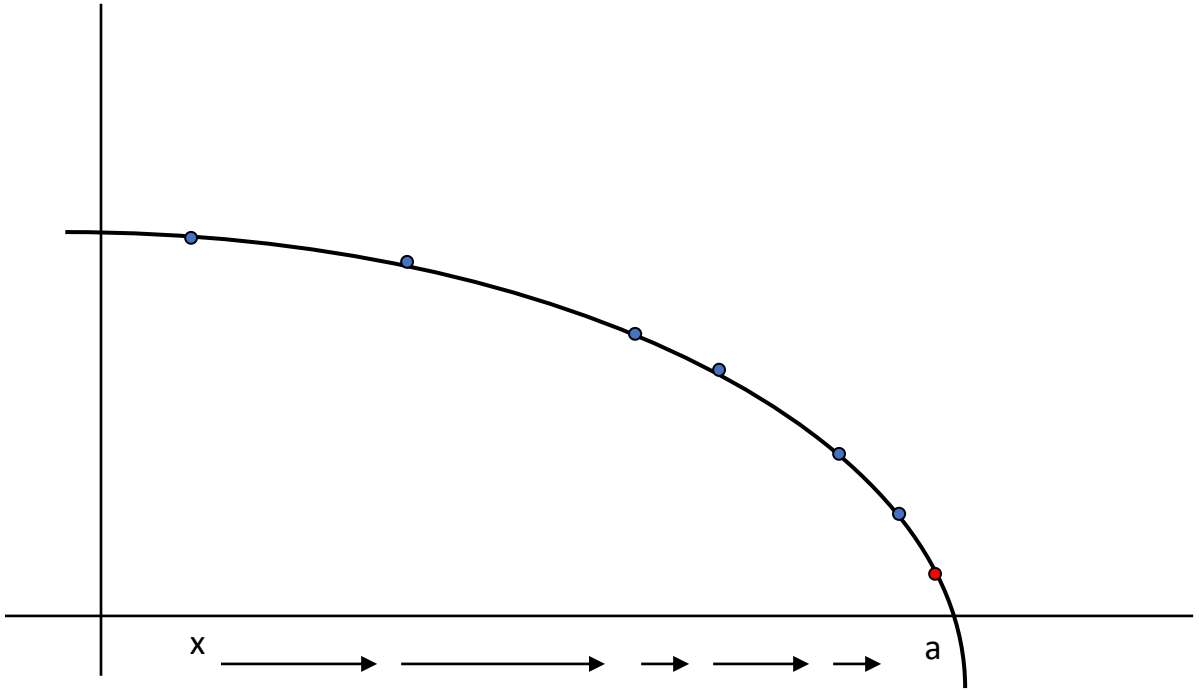
closer and closer…



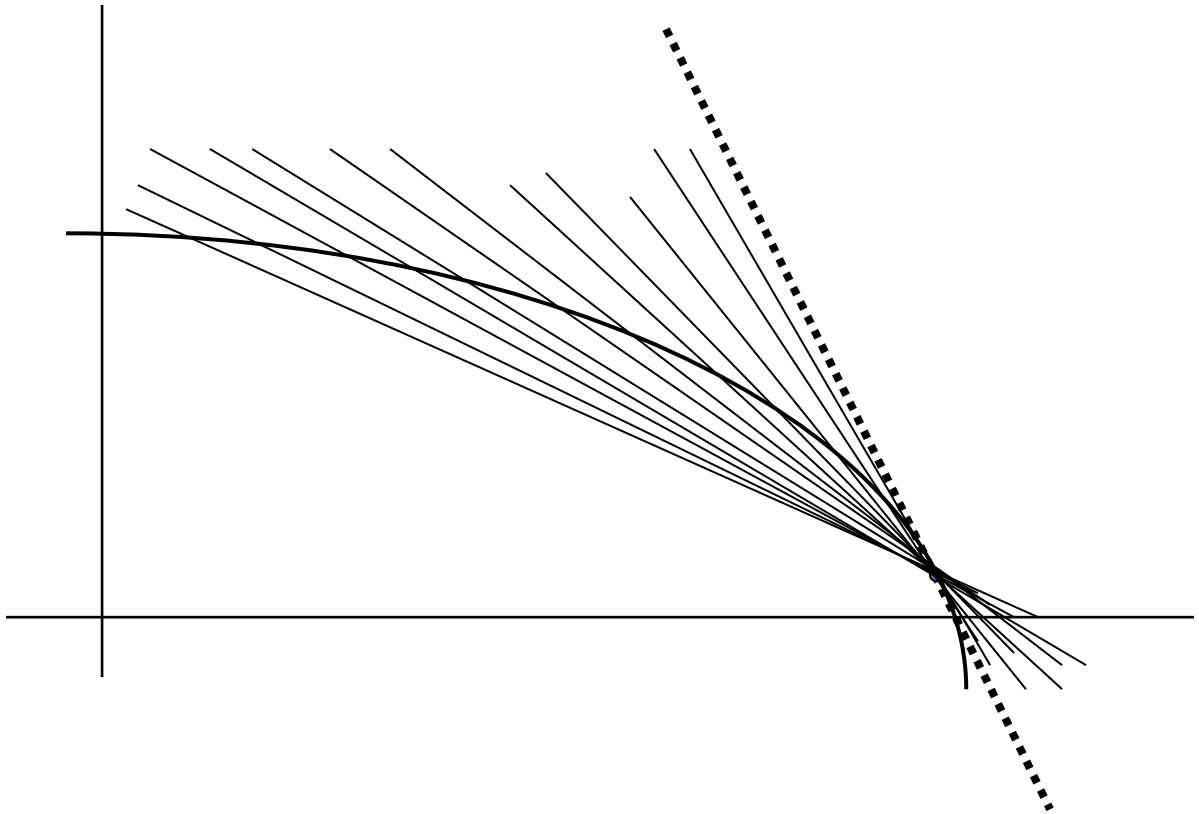
watch the slope...



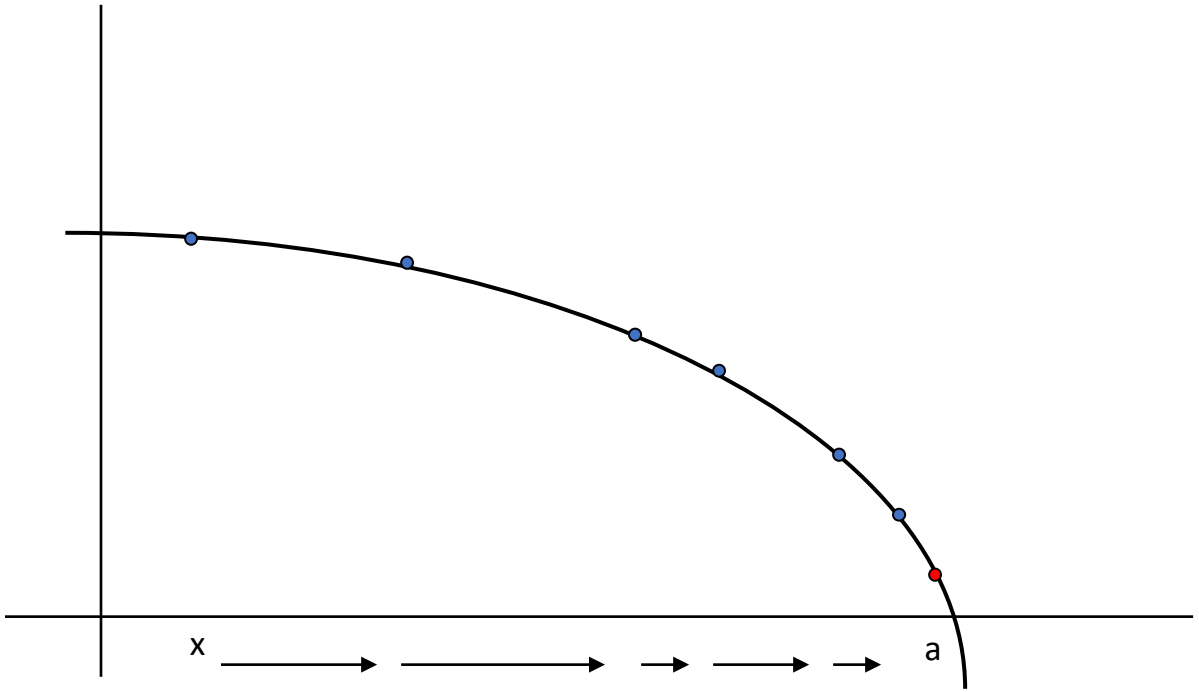
watch what *x* does...



The slope of the secant line gets closer and closer to the slope of the tangent line...



As the values of *x* get closer and closer to *a*!



The slope of the secant lines
gets closer
to the slope of the tangent line...

...as the values of x
get closer to a

Translates to….

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

as *x* goes to *a*

Equation for the slope

Which gives us the the exact slope
of the line tangent to the curve at *a*!

thus...

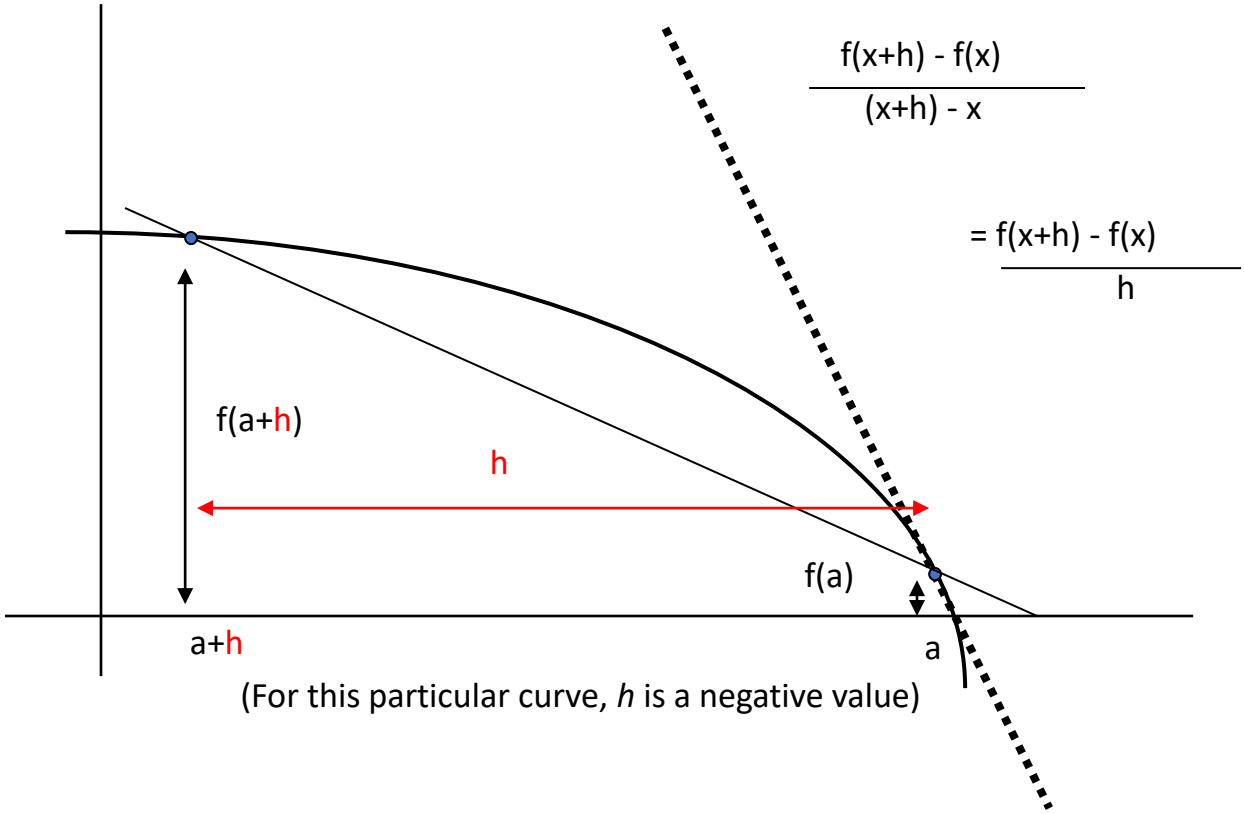
$$\lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

AND

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

Give us a way to calculate the slope of the line tangent at *a*!

similarly...

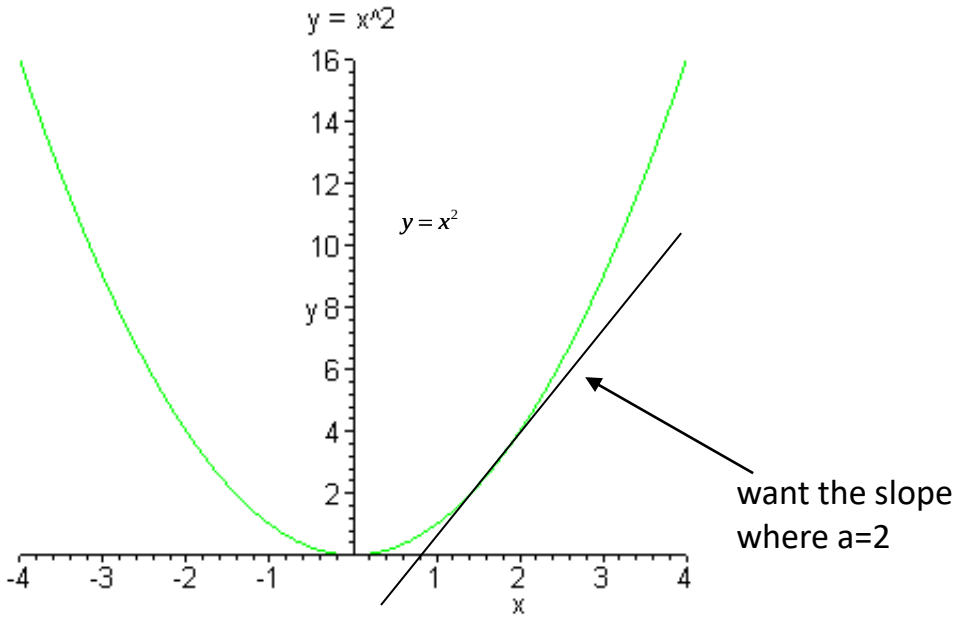

f(x+h) - f(x)
(x+h) - x
= f(x+h) - f(x)
h
f(a+h)
h
f(a)
a+h
a
(For this particular curve, h is a negative value)

Which one should I use?

(doesn't *really* matter)

A VERY simple example...

$$y = x^2$$



$$\lim \frac{f(x)-f(a)}{x-a} = \lim \frac{x^2-a^2}{x-a} = \lim \frac{(x-a)(x+a)}{x-a}$$

$$= \lim(x+a) = \lim(x+2) = 4$$

as x⟶
a=2

$$\lim \frac{f(x+h)-f(x)}{h} = \lim \frac{(x+h)^2 - x^2}{h}$$

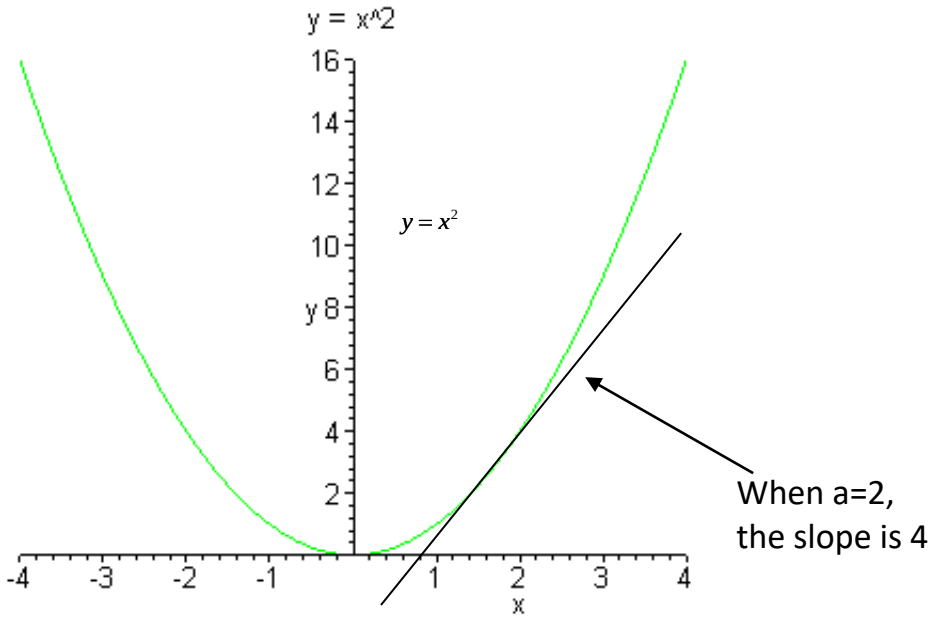
$$= \lim \frac{x^2 + 2xh + h^2 - x^2}{h} = \lim \frac{h(2x+h)}{h}$$

$$= \lim (2x+h) = 4$$

As h $\longrightarrow$ 0, x -> 2 in this example

back to our example...

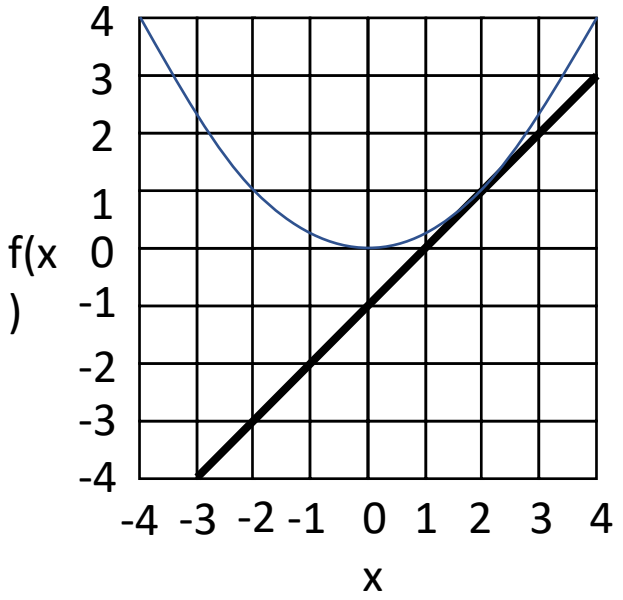
$$y = x^2$$



What is a derivative?

- The derivative $f'(x)$ of a function $f(x)$ says how fast $f(x)$ changes as x changes.
- Visually, $f'(x)$ is the slope of $f(x)$ at x.

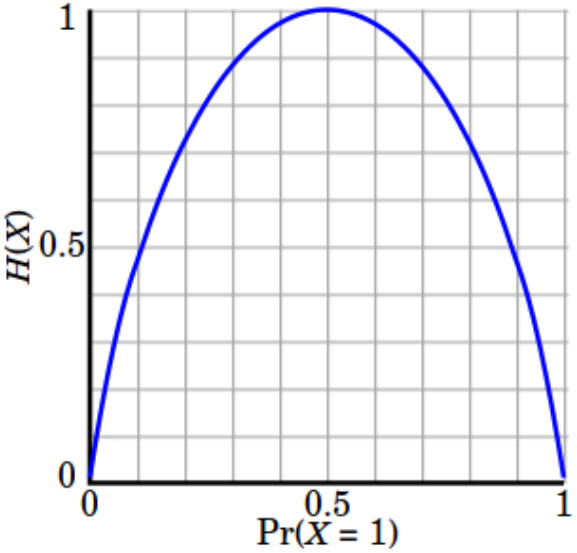
Example: If $f(x) = \frac{1}{4}x^2$ then $f'(2) = 1$ because the slope of $f(x)$ at $x = 2$ is 1. We can see this by looking at the tangent line to $f(x)$ at $x = 2$.



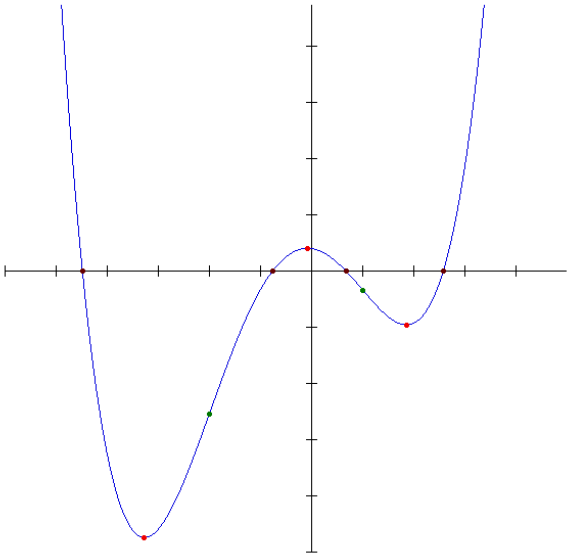
Why are derivatives useful?

- Tells us how quickly something is changing.
- In physics: velocity is the derivative of position and acceleration is the derivative of velocity (with respect to time).
- Optimization: Derivatives are crucial for finding the minimum or maximum of functions.
- And much much more.

How to Find the Maximum?



How To Find the Minimum?



Leibniz Notation

- So far, we have written the derivative of a function f as f'.
- Another notation, devised by Leibniz, is $\frac{df}{dx}$.
- Warning: $\frac{df}{dx}$ is a single function. df and dx do not have values on their own.

Advantages of Leibniz notation:

- Emphasizes how the derivative is computed. $\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}$
- Makes it easier to express the product rule, quotient rule, and chain rule

Disadvantage of Leibniz notation:

- Need clumsy notation like $\left(\frac{df}{dx}\right)_{x=2}$ or $\left.\frac{df}{dx}\right|_{x=2}$ to write the derivative of a function at a particular point.

in conclusion...

- The derivative is the the slope of the line tangent to the curve (evaluated at a point)
- it is a limit (2 ways to define it)
- once you learn the rules of derivatives, you WILL forget these limit definitions (beyond the scope of IST-718)
- cool site to go to for additional explanations http://archives.math.utk.edu/visual.calculus/2/

Probability and Statistics

What is Probability?

There are several possible interpretations of **probability**, but they (almost) completely agree on the mathematical rules that probability must follow.

- P(A) = Probability of event A
- $0 \leq P(A) \leq 1$

Random processes

- A ***random process*** is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples: coin tosses, die rolls, iTunes shuffle, whether the stock market goes up or down tomorrow, etc.
- It can be helpful to model a process as random even if it is not truly random.

MP3 Players › Stories › **iTunes: Just how random is random?**

iTunes: Just how random is random?

By David Braue on 08 March 2007

- Introduction
- Say You, Say What?
- A role for labels?
- The new random

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

http://www.cnet.com.au/itunes-just-how-random-is-random-339274094.htm

Statistical Learning

Two Views on Statistical Learning

- Bayesian statistics
- Frequentist Statistics

Bayesian

- A **Bayesian** interprets probability as a subjective degree of belief
- Provides an equation to manipulate conditional probabilities
- Uses *prior* knowledge about the phenomenon and then combine observed evidence with the prior knowledge.
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{likelihood * prior}{evidence}$
  - P(A): The prior knowledge
  - P(B): The posterior knowledge or evidence gained from additional observations
  - P(A|B): The posterior degree of belief after having observed B
  - P(B|A): Likelihood: The likelihood of A conditioned on B
  - The quotient $\frac{P(B|A)P(A)}{P(B)}$ represents the support B provides for A
- http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0809/eshky.pdf
- https://stats.stackexchange.com/questions/74082/what-is-the-difference-in-bayesian-estimate-and-maximum-likelihood-estimate

Frequentist

- **Frequentist statistics**:
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
  - Makes an estimate of a parameter based on sample data.
  - It assumes a procedure where samples from the same population will be observed an infinite amount of times.
  - It uses sampling with replacement
  - The samples build the distribution
  - We will use the frequentist interpretation in IST-718

Practice

Which of the following events would you be most surprised by?

(a) exactly 3 heads in 10 coin flips

(b) exactly 3 heads in 100 coin flips

(c) exactly 3 heads in 1000 coin flips

Practice

Which of the following events would you be most surprised by?

(a) exactly 3 heads in 10 coin flips

(b) exactly 3 heads in 100 coin flips

(c) exactly 3 heads in 1000 coin flips

Disjoint and non-disjoint outcomes

***Disjoint (mutually exclusive) outcomes*:**

Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

***Non-disjoint outcomes*:**

Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.

General Addition Rule

General addition rule

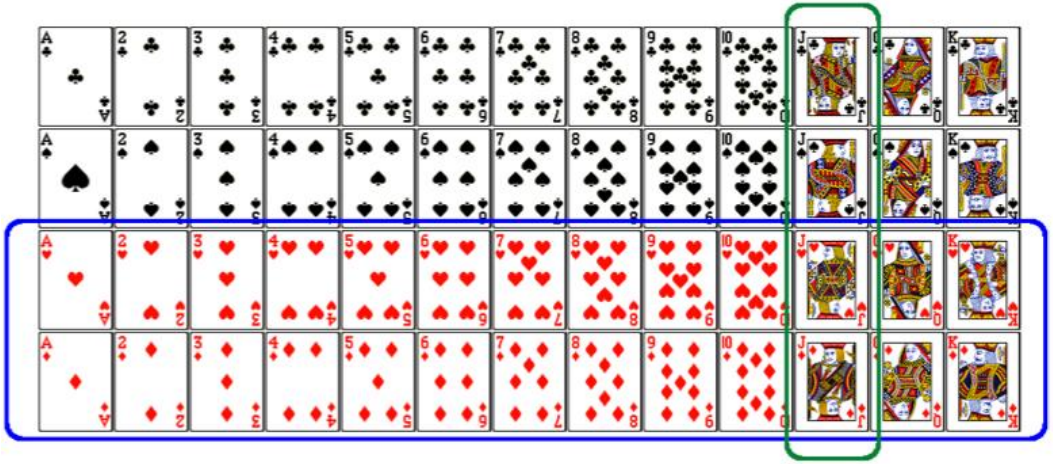
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Notes:

- For disjoint events $P(A \text{ and } B) = 0$, so the above formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$
- P(A or B) occurs means A, B, or both A and B

Union of non-disjoint events

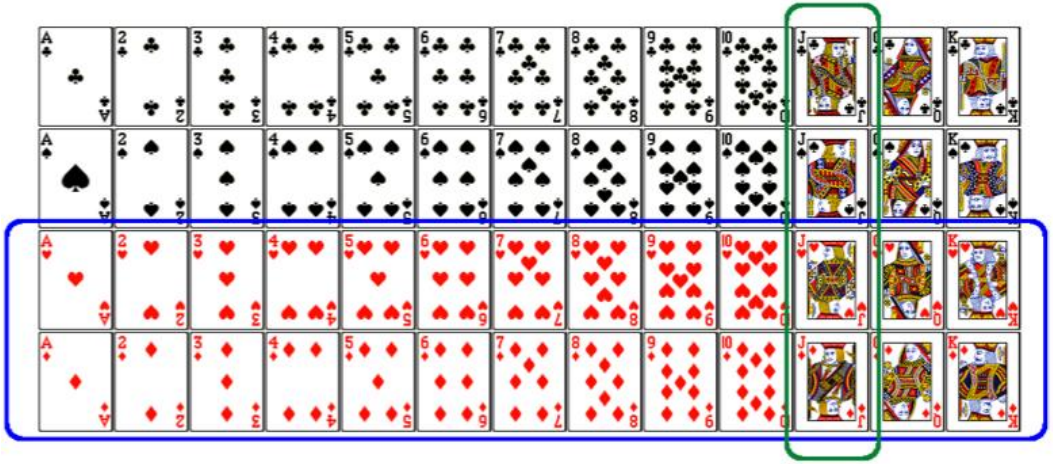
What is the probability of drawing a jack or a red card from a well shuffled full deck?



http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck?



$$P(jack\ or\ red) = P(jack) + P(red) - P(jack\ and\ red)$$

$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

http://www.milefoot.com/math/discrete/counting/cardfreq.htm

Practice

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

(a) (40 + 36 - 78) / 165

(b) (114 + 118 - 78) / 165

(c) 78 / 165

(d) 78 / 188

(e) 11 / 47

Practice

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

(a) (40 + 36 - 78) / 165

(b) (114 + 118 - 78) / 165

(c) 78 / 165

(d) 78 / 188

(e) 11 / 47

Probability distributions

A ***probability distribution*** lists all possible events and the probabilities with which they occur.

- The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

- Rules for probability distributions:
  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1

- The probability distribution for the genders of two kids:

Event	*MM*	*FF*	*MF*	*FM*
Probability	*0.25*	*0.25*	*0.25*	*0.25*

Random variables

A ***random variable*** is a numeric quantity whose value depends on the outcome of a random event

- We use a capital letter, like X, to denote a random variable
- The values of a random variable are denoted with a lowercase letter, in this case x
- For example, $P(X = x)$

There are two types of random variables:

- ***Discrete random variables*** often take only integer values
  - Example: Number of credit hours, Difference in number of credit hours this term vs last
- ***Continuous random variables*** take real (decimal) values
  - Example: Cost of books this term, Difference in cost of books this term vs last

Expectation and Expected Value

- We are often interested in the **average outcome** of a random variable.
- We call this the ***expected value*** (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

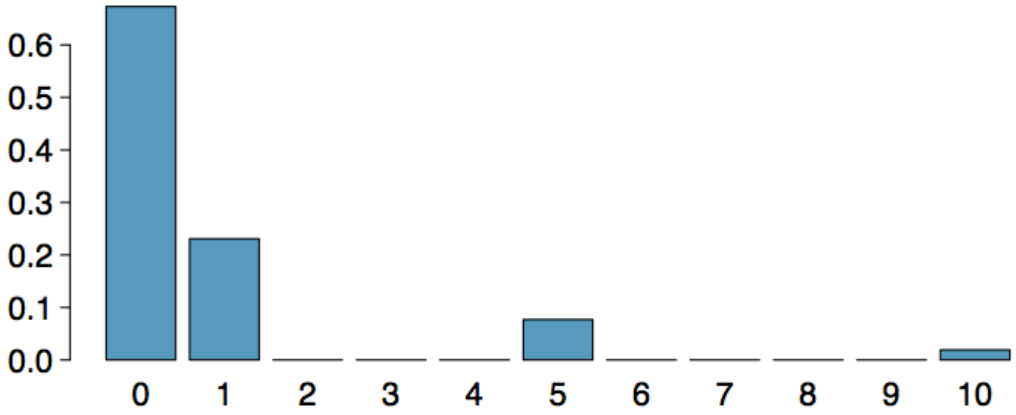
Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:

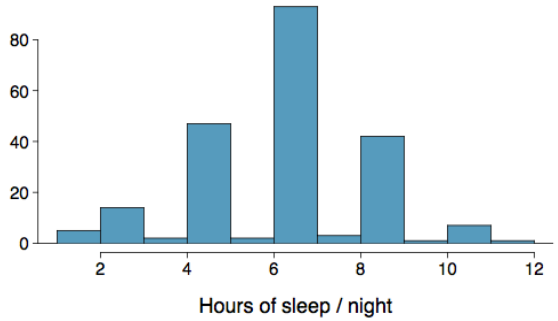


Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Calculate the variance for student sleep
- The sample mean is and the sample size is $\bar{x} = 6.71,$
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11\ hours^2$$

- Squared deviation rids of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

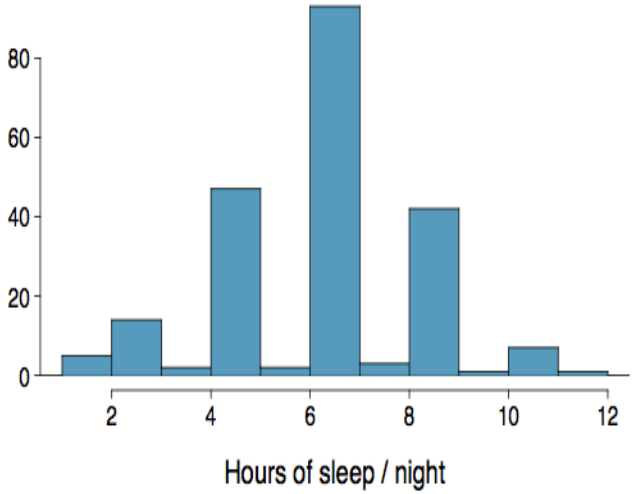
Standard Deviation

The ***standard deviation*** is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \; hours$$



- We can see that all of the data are within 3 standard deviations of the mean.

Marginal Probability

- The probability of an event occurring (p(A)), it may be thought of as an unconditional probability.
- Example: The probability of drawing a red card equals the number of red cards divided by the total number of cards: 26/52 = 0.5.

Sample Question: Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below. **What is the probability that a patient relapses?**

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed) = 48 / 72 ~ 0.67

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Joint probability

- (A and B). The probability of event A **and** event B occurring.
- The probability of the intersection of two or more events.
- The probability of the intersection of A and B may be written p(A ∩ B).
- Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

Sample Question: What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed and desipramine) = 10 / 72 ~ 0.14

Conditional probability

- p(A|B) is the probability of event A occurring, given that event B occurs.
- Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13.
- So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.
- The general formula for conditional probability follows

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse|desipramine) = \frac{P(relapse \text{ and } desipramine)}{P(desipramine)} = \frac{10/72}{24/72} = \frac{10}{24} = 0.42$$

Independent events

- If A and B represent two outcomes or events, then

  P(A and B) = P(A | B) x P(B)

  Note that this formula is simply the conditional probability formula, rearranged.
  - It is useful to think of A as the outcome of interest and B as the condition.
- Generically, if ***P(A | B) = P(A)*** then the events *A* and *B* are said to be **independent**.
- If two events are independent, their joint probability is simply the product of their probabilities.
  - Conceptually: Giving *B* doesn't tell us anything about *A*.
  - Mathematically: if events *A* and *B* are independent, *P(A and B) = P(A) x P(B)*. Then,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Example: independence and conditional probabilities

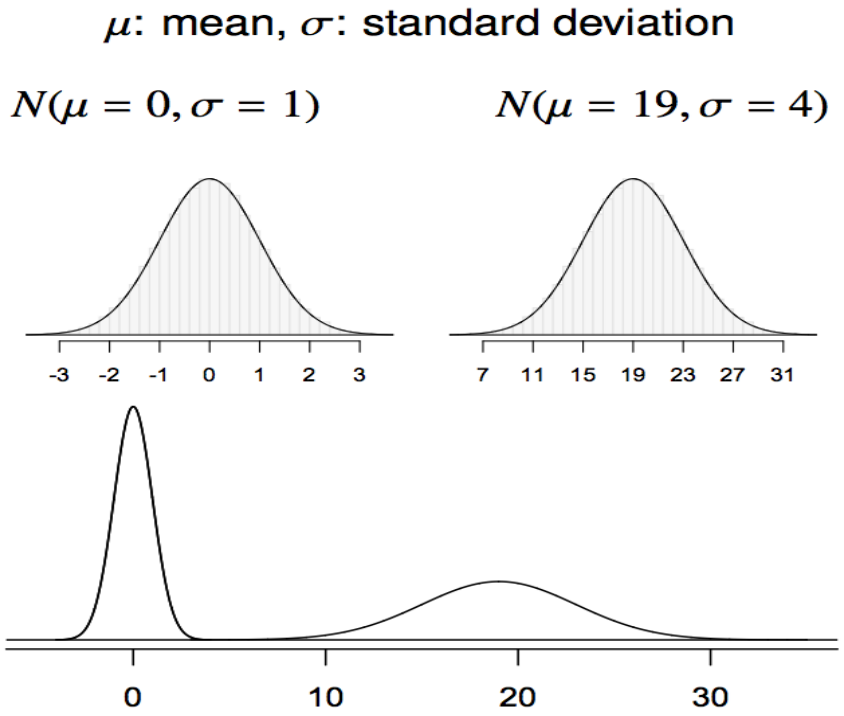
Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:

	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

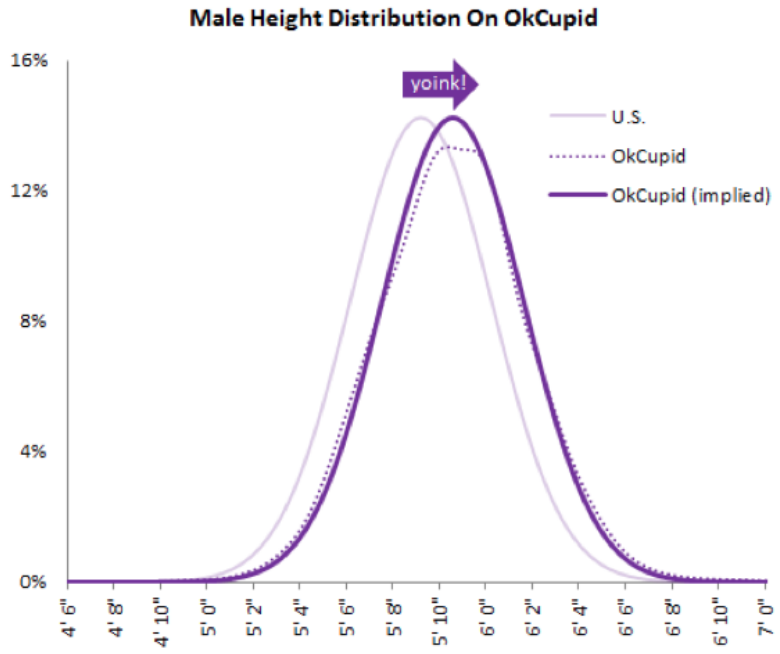
- The probability that a randomly selected student is a social science major is 60 / 100 = 0.6.
- The probability that a randomly selected student is a social science major given that they are female is 30 / 50 = 0.6.
- Since *P(SS | M)* also equals 0.6, major of students in this class does not depend on their gender: *P(SS | F) = P(SS)*.

Normal Distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as ***N(μ, σ)*** → Normal with mean μ and standard deviation σ



Normal Distribution: Heights of males

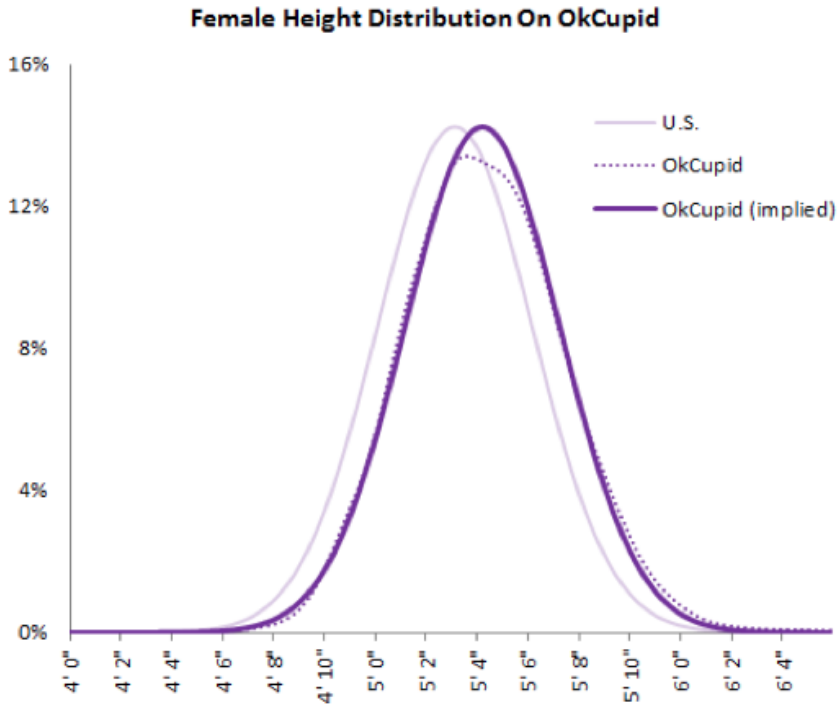


"The male heights on OkCupid very nearly follow the expected normal distribution -- except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating

Normal Distribution: Heights of females



"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

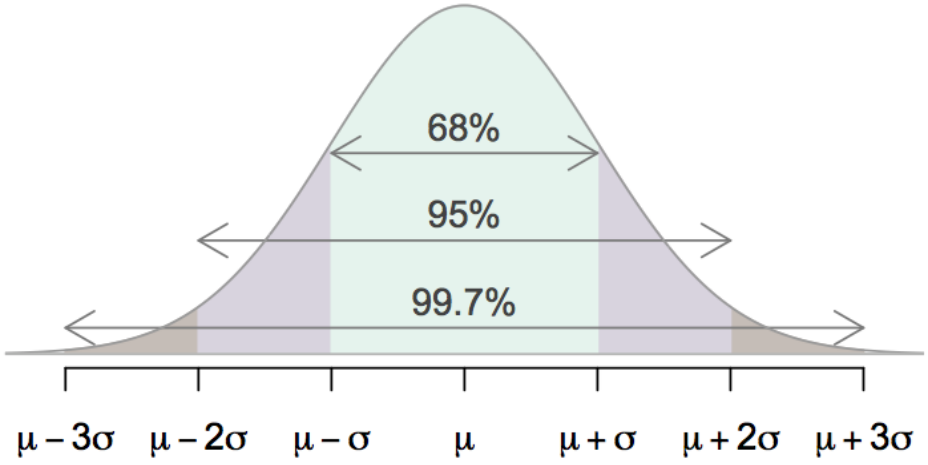
http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating

Normal Distribution 68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

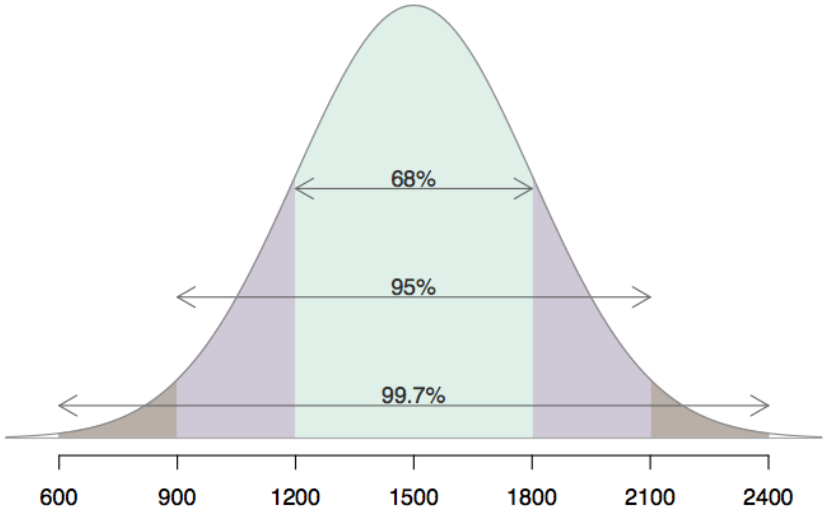
It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~$99.7% of students score between 600 and 2400 on the SAT.

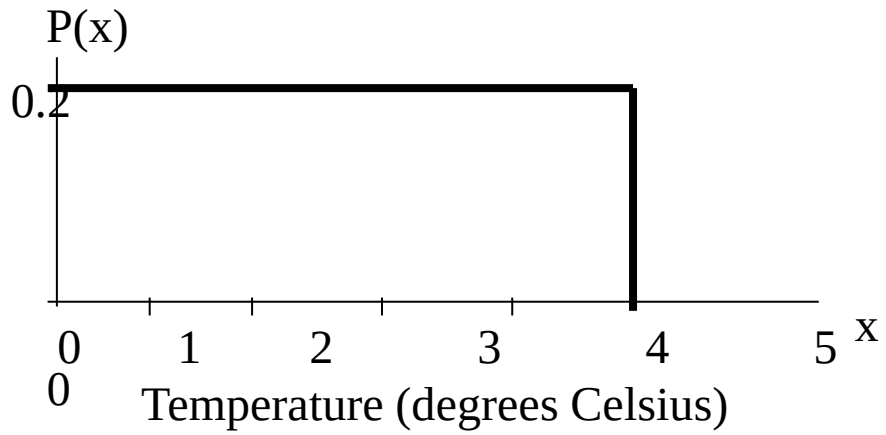


Uniform Distribution

A **Uniform Distribution** has equally likely values over range of possible outcomes.

A graph of the uniform probability distribution is a rectangle with area equal to 1.

Example: The figure below depicts the probability distribution for temperatures in a manufacturing process. The temperatures are controlled so that they range between 0 and 5 degrees Celsius, and every possible temperature is equally likely.



General Uniform Distribution

A **Uniform Distribution** has equally likely values over the range of possible outcomes, say c to d.

$$\text{Height of the density function}: f(x) = \frac{1}{d-c}$$

$$\text{Mean} = \mu = \frac{c+d}{2}$$

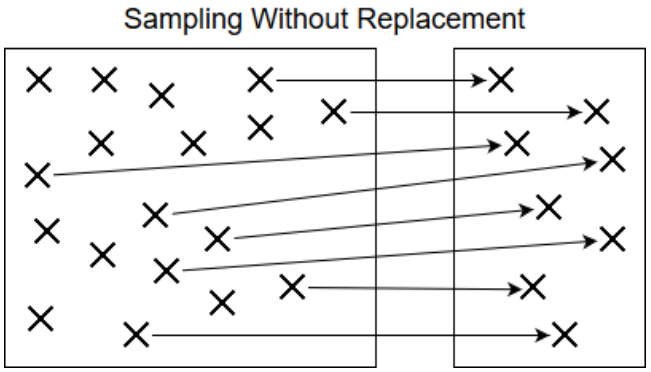
$$\text{Standard Deviation} = \sigma = \frac{d-c}{\sqrt{12}}$$

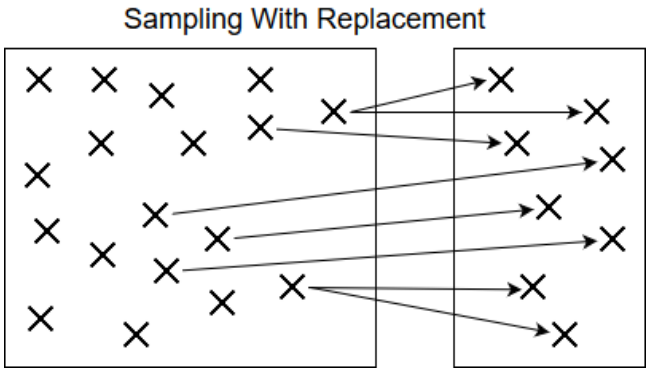
Population

- **Population**: A population is an entire collection of objects or individuals about which information is desired.
- Populations are often hard to measure exactly.
- If we want to measure the percentage of Trump voters in the United States, the population is all eligible voters in the United States.
- Some populations are known in their entirety.
- If we want to know the percentage of female employees at Lockheed in Liverpool, the population is all Lockheed Liverpool employees

Sampling

- What do we do if the population is too large to work with: Like the United states population?
- **Sample**: A sample is a subset of a population.
  - Ex: A sample of 1000 prospective voters for the Trump vs. Clinton election
  - Samples can be drawn from a population with or without replacement.
  - Sampling with or without replacement depends on context





Sampling Continued …

- The process of sampling is a natural process, we do it in our everyday lives …
- If you walked into an unfamiliar store and wanted to determine if the store was expensive or not, you would not check every price in the store; but rather, you would check the prices of a variety of items.
- If you are cooking a batch of spaghetti sauce and wanted to check the quality, you would taste a small sample instead of the entire batch.

Degrees of Freedom

- Degrees of freedom indicate the number of independent values that can vary when estimating a population parameter.
- Example:
  - Say you are trying to estimate a fixed population mean
  - You collect a sample containing 10 independent observations
  - Rearrange the formula for mean: $\mu * 10 = \sum_{i=1}^{10} x_i$ (or mean * 10 = sample sum)
  - Only 9 out of 10 values can vary; the 10^{th} sample value must be fixed
  - When estimating population parameters (like mean), the degrees of freedom = sample size – number of parameter estimates (10 – 1 in this example)

Bessel's Correction

- Bessel was a French Mathematician
- Bessel's Correction uses the notion of degrees of freedom to improve the estimate of population variance when estimated from sample data.
- Math Fact: Sample variance is always less than or equal to population variance (assuming sampling with replacement).
- Bessel's correction: Subtracting 1 from the number of sample observations in the variance formula produces a better estimate of the population variance.
- https://en.wikipedia.org/wiki/Bessel%27s_correction

Covariance

- Covariance measures the linear relationship between 2 random variables.
- Positive covariance means as X increases, Y also increases
- Negative covariance means as X increases, Y decreases
- Does have units, ranges from + infinity to - infinity
- Not normalized so it’s hard to tell the degree of co-variation based on the resulting number.
- The sign is what matters when interpreting covariance

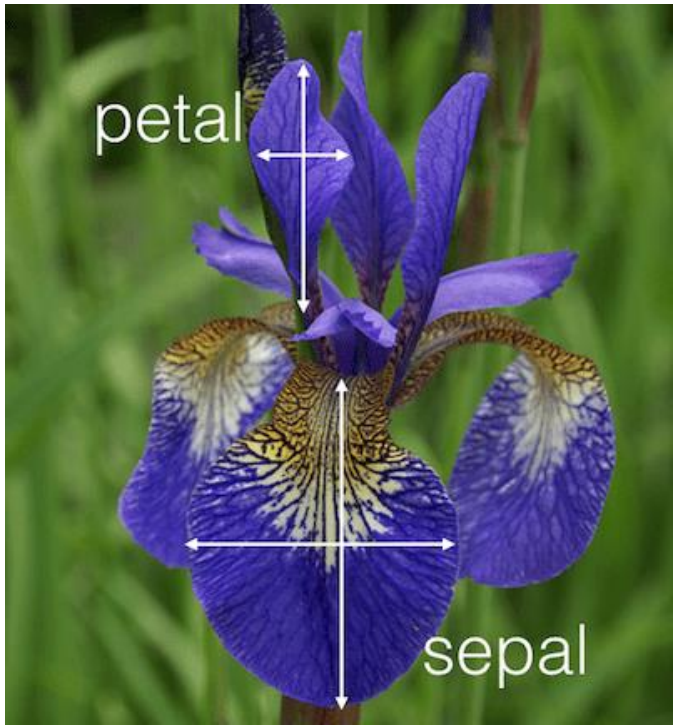
Covariance Continued …

- $COV(X,Y) = \frac{1}{n-ddof}(x_i - E(x))(y_i - E(y))$
- **numpy.cov(*m, y=None, rowvar=True, bias=False, ddof=None, fweights=None, aweights=None*)**
- ddof ==1, E(x) and E(y) are sample means
- ddof == 0, E(x) and E(y) are population means
- NumPy only has a function to calculate a covariance matrix, will not directly calculate covariance between 2 variables
- Note: Default is sample covariance (ddof == 1)
- Assumes that features in rows and observations in columns (rowvar == True)

Iris Data Covariance Matrix Example

- Iris Dataset comes from Statistician Ronald Fisher in 1936
- Covariance matrix for Iris data set.
- Cols: Sepal Length, Sepal Width, Petal Length and Petal Width
- Diagonal represents variance. For example, matrix[0][0] is sepal length variance

	0	1	2	3
0	0.685694	-0.0392685	1.27368	0.516904
1	-0.0392685	0.188004	-0.321713	-0.117981
2	1.27368	-0.321713	3.11318	1.29639
3	0.516904	-0.117981	1.29639	0.582414



Correlation Coefficient

- Correlation measures the linear relationship between 2 random variables.
- Dimensionless (no units)
- Ranges from -1 to 1 where the absolute correlation suggests the degree of correlation.
- Size and sign matters
- +1 is perfect positive correlation, -1 is perfect negative correlation, 0 is completely uncorrelated.

Correlation Continued

- Correlation is essentially a normalized covariance.
- $r = \frac{COV(X,Y)}{std_x * std_y}$
- std_x and std_y are the standard deviations of X and Y respectively
- **numpy.corrcoef(*x, y=None, rowvar=True, bias=<no value>, ddof=<no value>*)**
- Assumes features are in rows and observations are in columns.
- r = correlation coefficient
- Note: ddof is ignored in this command

Iris Data Correlation Matrix Example

- Correlation matrix for Iris dataset
- Cols: Sepal Length, Sepal Width, Petal Length and Petal Width
- Assumes that features in rows and observations in columns (rowvar == True)

	0	1	2	3
0	1	-0.109369	0.871754	0.817954
1	-0.109369	1	-0.420516	-0.356544
2	0.871754	-0.420516	1	0.962757
3	0.817954	-0.356544	0.962757	1

Correlation Rules of Thumb

Correlation Magnitude	Interpretation
0.00 – 0.20	Very Weak
0.20 – 0.40	Weak to Moderate
0.40 – 0.60	Medium to Substantial
0.60 – 0.80	Very Strong
0.80 – 1.00	Extremely Strong

Correlation / Covariance Summary

- Covariance provides the direction of a linear relationship between 2 variables
- Correlation provides the direction and strength of a linear relationship between 2 variables
- Covariance has units and ranges from negative infinity to positive infinity
- Correlation is unitless and ranges from -1 to +1
- Both can provide misleading results if provided with outliers or non linear data
- Both provide a measure of association, not causation

References

- https://www.deeplearningbook.org/slides/02_linear_algebra.pdf
- https://www.ling.upenn.edu/courses/cogs501/LinearAlgebraReview.html
- https://math.berkeley.edu/~xuwenzhu/classes/18.089/summer2015/Lecture/Lecture1.pptx
- http://www.math.ucla.edu/~tat/MicroTeach/deriv.ppt
- https://www.openintro.org/stat/teachers.php
- http://people.stfx.ca/jquinn/STAT%20231/Class%20Power%20points/Lecture%209%20Uniform%20and%20normal%20distributions.ppt
- https://sites.nicholas.duke.edu/statsreview/jmc/